import matplotlib.pyplot as plt

import seaborn as sns

import matplotlib as mpl

import matplotlib.pylab as pylab

import numpy as np

%matplotlib inline

import re

sentences = """ The speed of transmission is an important point of difference between the two viruses. Influenza has a shorter median incubation period (the time from infection to appearance of symptoms) and a shorter serial interval (the time between successive cases) than COVID-19 virus. The serial interval for COVID-19 virus is estimated to be 5-6 days, while for influenza virus, the serial interval is 3 days. This means that influenza can spread faster than COVID-19.

Further, transmission in the first 3-5 days of illness, or potentially pre-symptomatic transmission –transmission of the virus before the appearance of symptoms – is a major driver of transmission for influenza. In contrast, while we are learning that there are people who can shed COVID-19 virus 24-48 hours prior to symptom onset, at present, this does not appear to be a major driver of transmission.

The reproductive number – the number of secondary infections generated from one infected individual – is understood to be between 2 and 2.5 for COVID-19 virus, higher than for influenza. However, estimates for both COVID-19 and influenza viruses are very context and time-specific, making direct comparisons more difficult. """

sentences = re.sub('[^A-Za-z0-9]+', ' ', sentences) #remove the special char

sentences = re.sub(r'(?:^| )\w(?:$| )', ' ', sentences).strip() #remove 1 letter words

sentences = sentences.lower()

words = sentences.split()

vocab = set(words)

vocab\_size = len(vocab)

word\_to\_ix = {word: i for i, word in enumerate(vocab)}

ix\_to\_word = {i: word for i, word in enumerate(vocab)}

data = []

for i in range(2, len(words) - 2):

context = [words[i - 2], words[i - 1], words[i + 1], words[i + 2]]

target = words[i]

data.append((context, target))

print(data[:5])

embeddings = np.random.random\_sample((vocab\_size, 10))

def linear(m, theta):

w = theta

return m.dot(w)

def log\_softmax(x):

e\_x = np.exp(x - np.max(x))

return np.log(e\_x / e\_x.sum())

def NLLLoss(logs, targets):

out = logs[range(len(targets)), targets]

return -out.sum()/len(out)

def log\_softmax\_crossentropy\_with\_logits(logits,target):

out = np.zeros\_like(logits)

out[np.arange(len(logits)),target] = 1

softmax = np.exp(logits) / np.exp(logits).sum(axis=-1,keepdims=True)

return (- out + softmax) / logits.shape[0]

def forward(context\_idxs, theta):

m = embeddings[context\_idxs].reshape(1, -1)

n = linear(m, theta)

o = log\_softmax(n)

return m, n, o

def backward(preds, theta, target\_idxs):

m, n, o = preds

dlog = log\_softmax\_crossentropy\_with\_logits(n, target\_idxs)

dw = m.T.dot(dlog)

return dw

def optimize(theta, grad, lr=0.03):

theta -= grad \* lr

return theta

# c. Train model

theta = np.random.uniform(-1, 1, (2 \* 2 \* 10, vocab\_size))

epoch\_losses = {}

for epoch in range(80):

losses = []

for context, target in data:

context\_idxs = np.array([word\_to\_ix[w] for w in context])

preds = forward(context\_idxs, theta)

target\_idxs = np.array([word\_to\_ix[target]])

loss = NLLLoss(preds[-1], target\_idxs)

losses.append(loss)

grad = backward(preds, theta, target\_idxs)

theta = optimize(theta, grad, lr=0.03)

epoch\_losses[epoch] = losses

ix = np.arange(0,80)

fig = plt.figure()

fig.suptitle('Epoch/Losses', fontsize=20)

plt.plot(ix,[epoch\_losses[i][0] for i in ix])

plt.xlabel('Epochs', fontsize=12)

plt.ylabel('Losses', fontsize=12)

def predict(words):

context\_idxs = np.array([word\_to\_ix[w] for w in words])

preds = forward(context\_idxs, theta)

word = ix\_to\_word[np.argmax(preds[-1])]

return word

# (['we', 'are', 'to', 'study'], 'about')

#predict(['we', 'are', 'about', 'study'])

def accuracy():

wrong = 0

for context, target in data:

if(predict(context) != target):

wrong += 1

return (1 - (wrong / len(data)))

accuracy()

# d. Output

predict(['transmission', 'is', 'important', 'point'])